Risk-Sensitive Reward-Free Reinforcement Learning with CVaR

Xinyi Ni [1] **Guanlin Liu** [1] **Lifeng Lai** [1]

Abstract

Exploration is a crucial phase in reinforcement learning (RL). The reward-free RL paradigm, as proposed by (Jin et al., 2020), offers an efficient method to design exploration algorithms for risk-neutral RL across various reward functions with a single exploration phase. However, as RL applications in safety critical settings grow, there's an increasing need for risk-sensitive RL, which takes potential risks into consideration for decision-making. Yet, efficient exploration strategies for risk-sensitive RL remain underdeveloped. This study presents a novel risk-sensitive reward-free framework based on Conditional Value-at-Risk (CVaR), designed to effectively address CVaR RL for any given reward function through a single exploration phase. We introduce an efficient algorithm named CVaR-RF-UCRL, which is shown to be (ϵ, p)-PAC, with a sample complexity upper bounded by $\tilde{\mathcal{O}}\left(\frac{S^2AH^4}{\epsilon^2\tau^2}\right)$ with τ being the risk tolerance parameter. We also prove a $\Omega\left(\frac{S^2AH^2}{\epsilon^2\tau}\right)$ lower bound for any CVaR-RF exploration algorithm, demonstrating the near-optimality of our algorithm. Additionally, we propose the planning algorithms: CVaR-VI and its more practical variant, CVaR-VI-DISC. The effectiveness and practicality of our CVaR reward-free approach are further validated through numerical experiments.

1. Introduction

In reinforcement learning (RL), agents learn optimal actions by iteratively interacting with the environment and leveraging feedback from reward signals. A critical part of this learning process is *exploration*, where agents navigate through states to effectively gather environment information. Despite exploration being widely recognized as a vital aspect of RL, simple randomized exploration strategies often fail due to high sample complexity (Li, 2012). While research by (Dann & Brunskill, 2015; Dann et al., 2017; Azar et al., 2017; Jin et al., 2018) demonstrates that stochastic exploration can be sample-efficient, applying these algorithms across different reward functions can lead to inefficiencies. To address this, (Jin et al., 2020) introduces the concept of reward-free RL, in which the goal is to approximate the near optimal policy under any reward function after a single phase of exploration, enhancing the efficiency and adaptability of the learning process. (Jin et al., 2020) also derives upper and lower bounds of the sample complexity of the risk-free approach.

Building on these insights, subsequent studies such as (Wang et al., 2020; Zhang et al., 2021; Kaufmann et al., 2021; Ménard et al., 2021; Chen et al., 2022; Miryoosefi & Jin, 2022) have sought tighter upper bounds and more practical algorithms. The focus of these existing reward-free RL research has been predominantly on the risk-neutral approach, in which the goal is to maximize the average total (discounted) reward. In practical scenarios especially in those safety critical scenarios, however, decision-makers often have risk preferences that aim to mitigate low-probability but high-impact outcomes. Thus, the need to consider risks beyond solely optimizing for the average becomes important, leading to the development of risk-sensitive RL.

In risk-sensitive RL, the objective function is shaped by applying risk measures to reward functions (Delage & Mannor, 2010; Bäuerle & Ott, 2011; La & Ghavamzadeh, 2013; Shen et al., 2014; Fei et al., 2020; Prashanth et al., 2022; Ying et al., 2022), thus is also significantly dependent on the exploration phase. Various risk measures have been extensively analyzed and adopted in RL frameworks (Chow & Ghavamzadeh, 2014; Chow et al., 2015; Tamar et al., 2015b;a; Chow et al., 2017; Keramati et al., 2020; Ni & Lai, 2022a;b; Hau et al., 2023). One widely used class of risk measures is coherent risk measures, which satisfy a set of natural and desirable properties: *1) monotonicity, 2) translation invariance, 3) subadditivity, 4) positive homogeneity*, ensuring rationality and reliability in capturing risk preferences (Artzner et al., 1999; Tamar et al., 2015a). Despite these advancements, efficient exploration, particularly in contexts without a predefined reward function,

[1]Department of Electrical and Computer Engineering, University of California, Daivs, Davis, USA. Correspondence to: Xinyi Ni <xni@ucdavis.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

remains an under-explored area. Existing studies on sample complexity and algorithm performance in risk-sensitive RL typically target specific reward functions, potentially limiting their effectiveness in varied reward settings (Fei et al., 2021; Bastani et al., 2022; Du et al., 2022; Wang et al., 2023; Ding et al., 2023). This situation underscores the urgency for developing efficient exploration methods in risk-sensitive RL, crucial for its practical deployment and success in diverse stochastic environments. In this paper, we study risk-sensitive RL in the reward-free setting, and aim to answer the following question:

Is it possible to design provably efficient risk-sensitive reward-free RL algorithm?

In this paper, we design an algorithm with near-optimal sample complexity to the above question.

Contribution: This paper introduces a CVaR-based risk-sensitive reward-free RL framework (CVaR-RF RL). For the exploration phase, we propose CVaR-RF-UCRL to efficiently explore environments with unknown reward functions. The number of trajectories collected in the exploration phase is upper bounded by $\tilde{\mathcal{O}}\left(\frac{S^2AH^4}{\epsilon^2\tau^2}\right)$, where S is the number of states, A is the action count, H is the horizon length, ϵ is the targeted accuracy, and τ the risk tolerance level for CVaR. We also prove a lower bound of $\Omega\left(\frac{S^2AH^2}{\epsilon^2\tau}\right)$ for any CVaR-RF exploration algorithm. Subsequently, we introduce the CVaR-RF-planning algorithm equipped with CVaR-VI, which is able to solve CVaR RL for given reward function but without interacting with the environment. We also propose CVaR-VI-DISC, a discretized version of CVaR-VI for direct implementation in real-world settings while maintaining an optimization error within $\epsilon/3$. These developments ensure the efficiency and applicability of our CVaR-RF framework in advancing the field of risk-sensitive RL.

Challenges: 1). Compared to risk-neutral reward-free RL (Jin et al., 2020), CVaR-RF RL focuses only on the tail distribution related to the risk tolerance parameter τ. But in a reward-free setup, we can't access reward information, including the reward distribution. Therefore, we must adjust our exploration strategy based on τ. To address this, we define an adaptive stopping rule for different τ values during the exploration phase. Moreover, while the optimal policy in risk-neutral RL is Markovian, the optimal policy for risk-sensitive RL is history-dependent, which makes it more complex. To simplify this, we propose a planning algorithm with CVaR-VI that can construct a Markovian policy as the optimal policy for CVaR RL, reducing the added complexity. 2). Compared with CVaR RL (Chow & Ghavamzadeh, 2014; Chow et al., 2015; Tamar et al., 2015b;a; Chow et al., 2017; Keramati et al., 2020; Ni & Lai, 2022a;b; Hau et al., 2023), CVaR-RF RL faces challenges due to the absence of immediate feedback on risks associated with actions during the exploration phase. In CVaR RL, with rewards given, the agent doesn't need to explore every state or action, as it can immediately adjust its strategy based on the reward. However, in CVaR-RF RL, where rewards are unknown during the exploration, it's necessary to thoroughly explore the environment by visiting all possible states and actions. This extensive exploration gathers enough information for the planning phase, allowing the agent to adjust its strategy effectively. To facilitate this, we introduce CVaR-RF-UCRL, a method that efficiently explores all states.

Outline: In Section 2, we introduce the preliminaries essential for the understanding of CVaR-RF RL. Section 3 presents the formal problem statement of CVaR-RF RL. In Section 4, we present the CVaR-RF-UCRL for exploration and CVaR-RF-planning algorithms, and present the upper bound for sample complexity. Section 5 provides our analysis of the lower bound of sample complexity specifically for CVaR-RF exploration. Section 6 provides numerical examples. Section 7 offer concluding remarks.

2. Preliminaries

We explore a tabular Markov decision process (MDP) represented as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$. Here, $\mathcal{S}$ and $\mathcal{A}$ are state and action spaces with sizes S and A respectively, H is the number of steps per episode, $\mathbb{P}_h(\cdot|s,a)$ is the state transition probability at step h for action a in state s, and $r_h(s,a)$ is a deterministic reward function mapping state-action pairs to rewards between 0 and 1. Both transition probabilities and reward function can vary with each step h. We define $\Pi_{\mathcal{H}}$ as the set of history-dependent policies, where each policy π consists of H functions $\{\pi_h : \mathcal{S} \times \mathcal{H}_h \to \Delta_{\mathcal{A}}\}$. Here, $\mathcal{H}_h$ is the history up to step h, and $\Delta_{\mathcal{A}}$ is the probability simplex over $\mathcal{A}$. The probability of reaching state s under policy π is $P^\pi(s)$.

In each episode of the MDP, the process starts by choosing an initial state s_1 from an unknown initial distribution $\mathbb{P}_1(\cdot)$. At every step h, the agent observes the current state s_h from the state space $\mathcal{S}$, selects an action a_h based on the distribution $\pi_h(s_h; \mathcal{H}_h)$, earns a reward $r_h(s_h, a_h)$, and then moves to the next state s_{h+1} according to the transition probability $\mathbb{P}_h(\cdot|s_h, a_h)$. The episode ends when the agent reaches the state s_{H+1}.

We now introduce CVaR, a widely used coherent risk measure in RL (Rockafellar et al., 2000). For a random variable X, CVaR at a given risk tolerance $\tau \in (0, 1]$ is defined as:

$$\text{CVaR}_\tau(X) := \sup_{b \in \mathbb{R}} \left(b - \tau^{-1}\mathbb{E}[(b - X)^+]\right),$$

where the notation $x^+ = \max(x, 0)$. CVaR effectively quantifies the average outcome in the worst τ-percentile of scenarios. It is noteworthy that for a continuous variable X,

this definition aligns precisely with outcomes less than or equal to the τ-th quantile, as elucidated by (Acerbi & Tasche, 2002). This τ-th quantile is also identified as Value-at-Risk (VaR), another well-recognized risk measure. However, VaR lacks the coherence property, distinguishing it from CVaR.

It is important to note that the reward function in this context is deterministic, with its cumulative sum ranging between $[0, H]$. Given this constraint and acknowledging that the optimal b aligns with the VaR (see Lemma D.2), and considering $\text{VaR}_\tau \in [0, H]$, we can appropriately restrict the range of b as follows:

$$\text{CVaR}_\tau(X) := \sup_{b\in[0,H]} \left(b - \tau^{-1}\mathbb{E}[(b-X)^+]\right). \quad (1)$$

Reward-Free RL: The RF-RL framework, as proposed by (Jin et al., 2020), is structured into two distinct phases: exploration and planning. In the exploration phase, the goal is to design algorithms that can efficiently explore the environment without reward information. Formally, in the exploration phase, each episode commences with an exploration policy π^t, based solely on data from previous episodes. An episode ξ_t captures a sequence of states and actions $(s_1^t, a_1^t, \ldots, s_H^t, a_H^t)$, starting at initial state s_1^t. Actions are chosen as $a_h^t = \pi_h^t(s_h^t)$, with subsequent states determined as $s_h^t \sim \mathbb{P}_h(s_{h-1}^t, a_{h-1}^t)$. Each trajectory ξ_t is added to the dataset $\mathcal{D}_t$. Data collection ends at a random stopping time t_{stop}, resulting in dataset $D_{t_{\text{stop}}}$. Based on the dataset, we are able to get the empirical transition kernel $\hat{\mathbb{P}}$.

In the planning phase, the agent's exploration strategy is critically assessed. During this phase, the agent is no longer permitted to interact with the environment. Instead, a specific reward function r is given, and the primary goal is to derive a near-optimal policy tailored to this r using the dataset $D_{t_{\text{stop}}}$ gathered during the exploration phase. The efficiency of the exploration approach is quantified based on the number of trajectories needed to consistently reach this objective, effectively measuring the algorithm's ability to prepare the agent for diverse reward scenarios without direct interaction with the MDP.

Our Goal: This paper focuses on establishing an efficient CVaR based reward-free RL framework, including:
1). Develop a CVaR-RF-Exploration algorithm that efficiently explores the environment without requiring any reward function and is adaptive to different τ.
2). Propose a CVaR-RF-Planning algorithm, which computes near-optimal policies based on the dataset acquired during the exploration phase and a specified reward function, without further interaction with the environment.
3). Ensure the efficiency and reliability by analyzing the sample complexity of exploration algorithm and the optimization error of planning algorithm.

3. Problem Statement

To address the inner objective of CVaR outlined in (1), which depends on the variable b, we consider an augmented MDP, in which an augmented state is defined as $(s, b) \in \mathcal{S}^{\text{Aug}} := \mathcal{S} \times [0, H]$. The initial state for a given $b_1 \in [0, H]$ is set to (s_1, b_1). Then, for each timestep $h = 1, \ldots, H$, the agent selects action a_h based on policy π_h, and updates b_{h+1} to $b_h - r_h$.

For any history-dependent policy $\pi \in \Pi_{\mathcal{H}}$, timestep $h \in [H]$, state $s_h \in \mathcal{S}$, budget $b_h \in [0, H]$, and history $\mathcal{H}$, we define the value function as:

$$V_h^\pi(s_h, b_h; \mathcal{H}_h) = \mathbb{E}_\pi\left[\left(b_h - \sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'})\right)^+ \Big| s_h, \mathcal{H}_h\right].$$

The CVaR objective following policy π, starting at s_1, is then expressed as:

$$\text{CVaR}_\tau^\pi(s_1) = \max_{b_1\in[0,H]} \{b_1 - \tau^{-1} V_1^\pi(s_1, b_1)\},$$

and the optimal CVaR objective is formulated as:

$$\begin{aligned}\text{CVaR}_\tau^\star(s_1) &= \max_{\pi\in\Pi_{\mathcal{H}}} \max_{b_1\in[0,H]} \{b - \tau^{-1} V_1^\pi(s_1, b_1)\} \\ &= \max_{b_1\in[0,H]} \{b_1 - \tau^{-1} \min_{\pi\in\Pi_{\mathcal{H}}} V_1^\pi(s_1, b_1)\}.\end{aligned} \quad (2)$$

The work of (Bäuerle & Ott, 2011) significantly advances our understanding by establishing the existence of an optimal policy $\rho^\star : \mathcal{S}^{\text{Aug}} \to \mathcal{A}$, which is deterministic and Markovian within the augmented MDP, denoted by $\mathcal{S}^{\text{Aug}} = \mathcal{S} \times [0, H]$. With a starting point of $b_1 \in [0, H]$ and initial state (s_1, b_1), the process unfolds as follows: for each $h = 1, 2, \ldots, H$, the action a_h is determined as $\rho^\star(s_h, b_h)$, the reward r_h as $r_h(s_h, a_h)$, the next state s_{h+1} evolves according to $P_h^\star(s_h, a_h)$, and the budget b_{h+1} is updated to $b_h - r_h$. The additional state b_h effectively tracks the residual budget from b_1, serving as a comprehensive summary of historical decisions for the CVaR RL problem.

The adoption of deterministic Markovian policies simplifies the decision-making process in MDPs, directly associating states with actions, thereby facilitating implementation and analytical processes. Consequently, without loss of optimality, the optimization problem in (2) simplifies to:

$$\text{CVaR}_\tau^\star(s_1) = \max_{b_1\in[0,H]} \{b_1 - \tau^{-1} \min_{\rho\in\Pi^{\text{Aug}}} V_1^\rho(s_1, b_1)\}, \quad (3)$$

where Π^{Aug} is the class of deterministic Markovian policies.

We now introduce the function definitions and the Bellman equations for the augmented MDP proposed in (Bellemare et al., 2023; Wang et al., 2023). For any policy $\rho \in \Pi^{\text{Aug}}$,

we define:

$$V_h^\rho(s_h, b_h) = \mathbb{E}_\rho\left[\left(b_h - \sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'})\right)^+ \middle| s_h, b_h\right], \quad (4)$$

and

$$Q_h^\rho(s_h, b_h, a_h) = \mathbb{E}_\rho\left[\left(b_h - \sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'})\right)^+ \middle| s_h, b_h, a_h\right]. \quad (5)$$

For notation convenience, we introduce the following definition:

$$[\mathbb{P}_h V_{h+1}](s_h, b_h, a_h) = \mathbb{E}_{s_{h+1}\sim\mathbb{P}(\cdot|s_h,a_h)}[V_{h+1}(s_{h+1}, b_{h+1})].$$

These functions adhere to the following Bellman equations:

$$\begin{aligned} V_h^\rho(s_h, b_h) &= \mathbb{E}_{a_h\sim\rho_h(s_h,b_h)}\left[Q_h^\rho(s_h, b_h, a_h)\right], \\ Q_h^\rho(s_h, b_h, a_h) &= [\mathbb{P}_h V_{h+1}](s_h, b_h, a_h), \end{aligned} \quad (6)$$

where $b_{h+1} = b_h - r_h$ and $V_{H+1}^\rho(s, b) = b_1^+ := \max(0, b_1)$. Similarly, we define the optimal conditions as:

$$\begin{aligned} V_h^\star(s_h, b_h) &= \min_{a\in\mathcal{A}} Q_h^\star(s_h, a_h, b_h), \\ \rho_h^\star(s_h, b_h) &= \text{argmin}_{a\in\mathcal{A}} Q_h^\star(s_h, b_h, a_h)], \\ Q_h^\star(s_h, b_h, a_h) &= \left[\mathbb{P}_h V_{h+1}^\star\right](s_h, b_h, a_h), \end{aligned} \quad (7)$$

where $b_{h+1} = b_h - r_h$ and $V_{H+1}^\star(s, b) = b_1^+ = \max(0, b_1)$.

Here we introduce a key fact shown in (Wang et al., 2020), which shows the optimality of Π^{Aug}.

Theorem 3.1. *(Optimality) For any* $b_1 \in [0, 1]$, $V_1^\star(s_1, b_1) = V_1^{\rho^\star}(s_1, b_1) = \inf_{\pi\in\Pi_H} V_1^\pi(s_1, b_1)$.

Theorem 3.1 suggest that we could compute $V_1^\star$ and $\rho^\star$ using dynamic programming (DP) if the true transitions $\mathbb{P}$ were known, following the classical Value Iteration procedure in standard RL. By executing $\rho^\star$ starting from $(s_1, b_1^\star)$ where $b_1^\star := \arg\max_{b_1\in[0,H]}\{b_1 - \tau^{-1}V_1^\star(s_1, b_1)\}$, we can attain the optimal CVaR value.

Based on these arguments, the goal of our paper is to identify a probably approximately correct (PAC) algorithm for CVaR-RF RL, characterized by specific performance and accuracy parameters (ϵ, δ), which is defined as follows:

Definition 3.2. (PAC algorithm for CVaR-RF) A CVaR-RF exploration algorithm is (ϵ, δ)-PAC with a given risk tolerance τ if for any reward function r,

$$\mathbb{P}\left(\mathbb{E}_{s_1\sim\mathbb{P}_1}\left[\text{CVaR}_\tau^\star(s_1; r) - \text{CVaR}_\tau^{\hat{\rho}}(s_1; r)\right] \le \epsilon\right) \ge 1-\delta.$$

Here, $\text{CVaR}_\tau^\star(s_1; r)$ is derived by executing optimal policy $\rho^\star$ starting from $(s_1, b_1^\star)$ under the true transition probabilities $\mathbb{P}$ and the reward function r with optimal initial budget $b_1^\star := \arg\max_{b_1\in[0,H]}\{b_1 - \tau^{-1}V_1^\star(s_1, b_1; r)\}$. Conversely, $\text{CVaR}_\tau^{\hat{\rho}}(s_1; r)$ is derived by executing the output policy in the planning phase $\hat{\rho}$ starting from $(s_1, \hat{b}_1)$ under the empirical transition probabilities $\hat{\mathbb{P}}$ and the reward function r with the near optimal initial budget obtained in the planning phase.

4. Main Results

In this section, we first analyze the exploration phase by assuming the optimization error during the planning phase is bounded. Inspired by (Fiechter, 1994; Kaufmann et al., 2021), we propose the CVaR-RF-UCRL, which is an (ϵ, δ)-PAC algorithm for CVaR-RF exploration, with the sample complexity upper bounded by $\tilde{\mathcal{O}}(S^2AH^4/\epsilon^2\tau^2)$. Then, in the planning phase, we propose a CVaR-RF-planning algorithm, adopting CVaR-VI and CVaR-VI-DISC, which satisfy the optimization error assumption.

Notation: Consider $(s_h^i, a_h^i, s_{h+1}^i)$ as the state, action, and next state observed by an algorithm at step h of episode i. For any step $h \in [H]$ and any state-action pair $(s, a) \in \mathcal{S}\times\mathcal{A}$, we define $n_h^t(s, a) = \sum_{i=1}^t \mathbb{I}\{(s_h^i, a_h^i) = (s, a)\}$ as the count of visits to the state-action pair (s, a) at step h in the first t episodes, and $n_h^t(s, a, s') = \sum_{i=1}^t \mathbb{I}\{(s_h^i, a_h^i, s_{h+1}^i) = (s, a, s')\}$. The empirical transitions are defined as:

$$\hat{\mathbb{P}}_h^t(s'|s, a) = \begin{cases} \frac{n_h^t(s,a,s')}{n_h^t(s,a)}, & \text{if } n_h^t(s, a) > 0 \\ \frac{1}{S}, & \text{otherwise} \end{cases}.$$

We denote by $\hat{V}_h^{t,\rho}(s_h, b_h; r)$ and $\hat{Q}_h^{t,\rho}(s_h, b_h, a_h; r)$ the value and action-value functions of a policy π in the MDP with transition kernels $\hat{\mathbb{P}}_t$ and reward function r.

4.1. Exploration Phase

We first present a lemma that will be useful for the study of the objective within the CVaR-RF exploration context. Prior to delving into this lemma, we make an assumption regarding the planning phase.

Assumption 4.1. The optimization error during the planning phase is bounded, i.e.,

$$\left|\widehat{\text{CVaR}}_\tau^{\hat{\rho}^\star}(s_1; r) - \widehat{\text{CVaR}}_\tau^{\hat{\rho}}(s_1; r)\right| \le \epsilon\tau/3,$$

where $\widehat{\text{CVaR}}_\tau^{\hat{\rho}^\star}(s_1; r)$ is derived by executing the optimal policy $\hat{\rho}^\star$ starting from $(s_1, \hat{b}_1^\star)$ under the empirical transition probabilities $\hat{\mathbb{P}}$ and the reward function r with optimal initial budget $\hat{b}_1^\star := \arg\max_{b_1\in[0,H]}\{b_1 - \tau^{-1}\hat{V}_1^\star(s_1, b_1; r)\}$.

Notice that, Assumption 4.1 focuses on the optimization error based on same emprical transition probability $\hat{\mathbb{P}}$ and given r. This assumption is not about the error with respect to the optimal policy for the true underlying MDP. Theorem 3.1 justifies the existence of an optimal policy $\hat{\rho}^*$ for MDP specified by $\hat{\mathbb{P}}$ and given reward function (more technical details could be found in Appendix A.1). Furthermore, there exist many CVaR RL works capable of generating such a near-optimal policy $\hat{\rho}$ that satisfies this assumption, such as (Chow et al., 2015; Tamar et al., 2015b; Wang et al., 2023). We also propose our algorithms in the planning phase that satisfy this assumption. Therefore, Assumption 4.1 could be immediately satisfied based on these facts.

The following lemma is useful for subsequent discussions and analyses related to our primary objective.

Lemma 4.2. *An algorithm is (ϵ, δ)-PAC for CVaR-RF exploration with a given risk tolerance τ if for any reward function r and for any $b_1 \in [0, H]$, $\left|V_1^{\rho}(s_1, b_1; r) - \hat{V}_1^{\rho}(s_1, b_1; r)\right| \leq \epsilon\tau/3$.*

Proof. Please refer to Appendix A.1 for more details. □

For simplifying the exposition of our algorithm, we posit that the initial state distribution $\mathbb{P}_0$ is supported solely on a singular state s_1. This assumption incurs no loss of generality (Fiechter, 1994). If this is not the case, one might contemplate an augmented MDP that includes an additional initial state s_0, paired with a unique action a_0 yielding a null reward. Thus, $b_0 = b_1$. In this scenario, the transitions from s_0 using a_0 are defined as $\mathbb{P}_0(\cdot|s_0, a_0) = \mathbb{P}_0$.

The error upper bounds in the CVaR-RF-UCRL algorithm are derived from an upper bound on the estimation error for each policy ρ, each initial budget $b \in [0, H]$ and each reward function r. The complete procedure is outlined in Algorithm 1. Before discussing the details of this algorithm, we introduce the definitions for the estimation error and its upper confidence bound.

Definition 4.3. For a given policy ρ, reward function r, and episode t, we define this error for any $(s_h, b_h, a_h) \in \mathcal{S}^{\text{Aug}} \times \mathcal{A}$ as

$$\hat{e}_h^{t,\rho}(s_h, b_h, a_h; r) := \left|\hat{Q}_h^{t,\rho}(s_h, b_h, a_h; r) - Q_h^{\rho}(s_h, b_h, a_h; r)\right|.$$

Definition 4.4. The upper confidence bound $E_h^t(s_h, a_h)$ for the error, recursively defined as follows: $E_{H+1}^t(s, a) = 0$ for all (s, a), and for all $h \in [H]$, with the convention $\frac{1}{0} = +\infty$,

$$E_h^t(s_h, a_h) = \min\left\{H, H\sqrt{\frac{2\beta(n_h^t(s,a), \delta)}{n_h^t(s,a)}} + \sum_{s'} \hat{\mathbb{P}}_h^t(s'|s,a) \max_a E_{h+1}^t(s', a)\right\}, \tag{8}$$

where $\beta(n, \delta)$ is a threshold function, an input to the algorithm, the choice of which will be discussed later.

It is important to note that the error upper bound only depends on the state s and action a, and is independent of the policy ρ, initial budget b_1 and reward function r. Lemma 4.5 establishes that $E_h^t(s, a)$ serves as the upper bound on the error $\hat{e}_h^{t,\rho}(s, b, a; r)$ for any ρ, b, r with a high probability.

Lemma 4.5. *With the Kullback-Leibler divergence between two distributions over $\mathcal{S}$ defined as $KL(p \parallel q) = \sum_{s\in\mathcal{S}} p(s) \log \frac{p(s)}{q(s)}$, consider the event*

$$\mathcal{E} = \left\{\forall t \in \mathbb{N}, \forall h \in [H], \forall (s, a), KL(\hat{\mathbb{P}}_h^t(\cdot|s,a), \mathbb{P}^h(\cdot|s,a)) \leq \frac{\beta(n_h^t(s,a), \delta)}{n_h^t(s,a)}\right\},$$

it is established that for any policy ρ, any reward function r and any b, $\hat{e}_h^{t,\rho}(s, b, a; r) \leq E_h^t(s, a)$ holds on event $\mathcal{E}$.

Proof. Please refer to the Appendix A.2 for more details. □

We now introduce the proposed CVaR-RF-UCRL algorithm, which operates on the principle of uniformly reducing the estimation error across all budgets, policies and potential reward functions by adopting a greedy approach based on the upper bounds E^t on these errors. The stopping criterion for CVaR-RF-UCRL is defined as reaching an error in step 1 that is smaller than $\epsilon\tau/3$:

- **Sampling rule:** the exploration policy π^{t+1} is the greedy policy with respect to $E^t(s, a)$, such that for all $s \in \mathcal{S}$ and $h \in [H]$:
$$\pi_h^{t+1}(s_h) = \text{argmax}_a E_h^t(s, a). \tag{9}$$
- **Stopping rule:** the algorithm stops at
$$t_{\text{stop}} = \inf\{t : E_h^t(s_1, \pi_1^{t+1}(s_1)) \leq \epsilon\tau/3\}.$$

Now, we have the following Lemma showing that CVaR-RF-UCRL is an algorithm with (ϵ, δ)-PAC.

Lemma 4.6. *(Correctness) On the event $\mathcal{E}$, given τ, for any r ,ρ and b_1, $\left|V_1^{t_{stop},\rho}(s_1, b_1; r) - \hat{V}_1^{t_{stop},\rho}(s_1, b_1; r)\right| \leq \epsilon\tau/3$, which implies $CVaR_{\tau}^{\star}(s_1; r) - CVaR_{\tau}^{\hat{\rho}^{\star}}(s_1; r) \leq \epsilon$.*

Algorithm 1 CVaR-RF-UCRL

Given: risk tolerance $\tau \in (0, 1]$
Initialization: $t = 1$, $\mathcal{D}_0 = \emptyset$, initialize E^0 with (8) and π_h^1 with (9)
while $E_h^{t-1}(s_1, \pi_1^t(s_1)) \geq \epsilon\tau/3$ **do**
  Observe the initial state $s_1^t \sim P_0$
  for $h = 1, \ldots, H-1, H$ **do**
    Play $a_h^t \sim \pi_h^t(s_h^t)$
    Observe the next state s_{h+1}
  end for
  Compute E^t according to (8) and π^{t+1} according to (9)
  $D_t = D_{t-1} \cup (s_1^t, a_1^t, \ldots, s_H^t, a_H^t)$
  $t = t + 1$
end while
Return the dataset $\mathcal{D}_{t_{\text{stop}}}$

Proof. By definition of the stopping rule and the sampling rule, we have for all $a \in \mathcal{A}$, $E_1^{t_{\text{stop}}}(s_1, a) \leq \epsilon/3$. Hence, by Lemma 4.5 on the event $\mathcal{E}$, for all ρ, b_1, r, and all a, $\hat{e}_1^{t_{\text{stop}},\rho}(s_1, b_1, a; r) \leq \epsilon\tau/3$. In particular, for all ρ, b_1, and r, $\left|V_1^{t_{\text{stop}},\rho}(s_1, b_1; r) - \hat{V}_1^{t_{\text{stop}},\rho}(s_1, b_1; r)\right| \leq \epsilon\tau/3$. The conclusion follows from Lemma 4.2 by choosing ρ to be $\hat{\rho}^\star$. □

We are now able to state our main results for CVaR-RF-UCRL, which show that with a proper chosen threshold $\beta(n, \delta)$, CVaR-RF-UCRL achieves (ϵ, δ)-PAC for CVaR RL. Furthermore, an upper bound on its sample complexity can be established under these conditions.

Theorem 4.7. *(Upper Bound for Sample Complexity) Using threshold* $\beta(n, \delta) = \log(2SAH/\delta) + (S-1)\log(e(1 + n/(S-1)))$*, the CVaR-RF-UCRL is* (ϵ, δ)*-PAC for CVaR-RF exploration. The number of trajectories collected in the exploration phase is bounded by* $\tilde{\mathcal{O}}\left(\frac{S^2AH^4}{\epsilon^2\tau^2}\right)$.

Proof. Please refer to Appendix A.3 for more details. □

Compared with the risk-neutral reward-free approaches (Jin et al., 2020; Kaufmann et al., 2021; Ménard et al., 2021), the denominator of the bound we obtained is related to the risk tolerance parameter τ. This is expected since CVaR is interpreted as the mean of the tail distribution with an area under the curve equal to τ, it requires more trajectories for smaller τ values and fewer trajectories for larger τ values.

4.2. Planning Phase

In the planning phase, the reward function is provided, and the goal is to find a near-optimal policy based on the given reward function and the dataset generated during the exploration phase. Following a similar approach to (Jin et al., 2020), we now introduce our planning algorithm, as outlined in Algorithm 2.

Algorithm 2 CVaR-RF-Planning

Input: a dataset of transition $\mathcal{D}_{t_{\text{stop}}}$, reward function r, accuracy ϵ, risk tolerance τ.
for all $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$ **do**
  $N_h(s, a, s') \leftarrow \sum_{(s_h, a_h, s_{h+1}) \in \mathcal{D}} \mathbb{I}[s_h = s, a_h = a, s_{h+1} = s']$.
  $N_h(s, a) \leftarrow \sum_{s'} N_h(s, a, s')$.
  $\hat{\mathbb{P}}_h(s'|s, a) = N_h(s, a, s')/N_h(s, a)$.
end for
$\hat{\rho}, \hat{b} \leftarrow$ APPROXIMATE-CVaR-SOLVER$(\hat{\mathbb{P}}, r, \epsilon, \tau)$.
return policy $\hat{\rho}$, and initial budget $\hat{b}$.

In Algorithm 2 , we first compute the empirical transition matrix $\hat{\mathbb{P}}$ based on the collected dataset $\mathcal{D}_{t_{\text{stop}}}$. Then, for each reward function r, we find a near-optimal policy by employing a APPROXIMATE-CVaR-SOLVER that utilizes transitions $\hat{\mathbb{P}}$, the given reward function r, an accuracy parameter ϵ and the given risk tolerance τ. It's worth noting that the solver can be any algorithm designed to find an $\epsilon/3$-suboptimal policy $\hat{\rho}$ for CVaR RL when both the transition matrix and the reward are known. One straightforward approach to achieve this is by using the Value Iteration algorithm, which iteratively solves the Bellman optimality equation (6) in a dynamic programming manner. The greedy policy induced by the resulting $Q^\star$ yields the optimal optimal policy without errors. We present Algorithm 3, which generates an optimal policy exactly according to Theorem 3.1 (Wang et al., 2023). This algorithm satisfies our Assumption 4.1 about the optimization error.

Algorithm 3 CVaR-VI

Input: transition matrix $\mathbb{P}$, reward function r, risk tolerance τ
for all $s \in \mathcal{S}, b \in [0, H]$ **do**
  Set $V_{H+1}(s, b) = b^+$
  for $h = H, H-1, \ldots, 1$ **do**
    $Q_h(s_h, b_h, a_h) = [\mathbb{P}_h V_{h+1}](s_h, b_h, a_h)$, where $b_{h+1} = b_h - r_h$
    $\rho_h^\star(s_h, b_h) = \text{argmin}_a Q_h(s_h, b_h, a_h)$
    $V_h^\star(s_h, b_h) = \min_a Q_h(s_h, b_h, a_h)$
  end for
end for
Calculate $b^\star = \text{argmax}_{b_1 \in [0,1]} \left\{b - \tau^{-1} V_1(s_1, b)\right\}$
return policy $\rho^\star$ and $b^\star$

4.2.1. Discretization

Algorithm 3 faces computational challenges due to the dynamic programming step, which requires optimization over

all $b \in [0, H]$, involving the maximization of a non-concave function (Wang et al., 2023). Following the approach proposed in (Bastani et al., 2022; Wang et al., 2023), we introduce a discretization of rewards, which allows the mentioned steps to be performed over a finite grid. This offers computational efficiency while preserving the same statistical guarantees.

We fix a precision $\eta \in (0, 1)$ and define $\phi(r) = \eta\lceil r/\eta \rceil \wedge 1$. This rounding function maps $r \in [0, 1]$ to an η-net of the interval $[0, 1]$, commonly referred as the grid. The discretized MDP dis($\mathcal{M}$) is an exact replica of the true MDP $\mathcal{M}$ with one exception: its rewards are post-processed using ϕ, i.e., $r(s, a; \text{disc}(\mathcal{M})) = \phi(r(s, a; \mathcal{M}))$. We now introduce the CVaR value iteration with discretization algorithm.

Algorithm 4 CVaR-VI-DISC

1: **Input:** transition matrix $\mathbb{P}$, reward function r, precision parameter η, risk tolerance τ.
2: Discretize the reward funtion r by

$$\hat{r} = \phi(r) = \eta\lceil r/\eta \rceil \wedge 1$$

3: **for** all $s \in \mathcal{S}, \hat{b} = n \cdot \eta, n = 0, 1, \ldots$ **do**
4: Set $\hat{V}_{H+1}(s, \hat{b}) = \hat{b}^+ := \max(0, \hat{b})$
5: **for** $h = H, H-1, \ldots, 1$ **do**
6: $\hat{Q}_h(s_h, \hat{b}_h, a_h) = \left[\mathbb{P}_h \hat{V}_{h+1}\right](s_h, \hat{b}_h, a_h)$, where $\hat{b}_{h+1} = \hat{b}_h - \hat{r}_h$
7: $\hat{\rho}^\star_h(s_h, \hat{b}_h) = \text{argmin}_a \hat{Q}_h(s_h, \hat{b}_h, a_h)$
8: $\hat{V}^\star_h(s_h, \hat{b}_h) = \min_a \hat{Q}_h(s_h, \hat{b}_h, a_h)$
9: **end for**
10: **end for**
11: Calculate $\hat{b}^\star = \text{argmax}_{\hat{b}} \left\{\hat{b} - \tau^{-1}\hat{V}_1(s_1, \hat{b})\right\}$
12: **return** policy $\hat{\rho}^\star$ and $\hat{b}^\star$

4.2.2. Computational Complexity

In disc($\mathcal{M}$), the τ-th quantile of the returns distribution (the argmax of the CVaR objective) will be a multiple of η. Therefore, it suffices to compute $V_1(s_1, b_1)$ and maximize line 9 over the grid. Since b_1 transitions by subtracting rewards, which are multiples of η, b_h will always stay on the grid. Hence, the entire dynamic programming procedure only needs to occur on the grid. This approach demonstrates that CVaR value iteration via discretization is computationally tractable.

Theorem 4.8. *The CVaR-VI-DISC has a run time of $\mathcal{O}(S^2AH\eta^{-2})$ in the discretized MDP. Setting $\eta = \epsilon\tau/3H$, as suggested in Theorem 4.9, the run time is $\mathcal{O}(\frac{S^2AH^3}{\epsilon^2\tau^2})$.*

Proof. Please refer to Appendix for more details. □

4.2.3. Discretization Error

Next, we evaluate the impact of errors resulting from the discretization step. Following a similar method as previous works (Wang et al., 2023), we can relate the errors within disc($\mathcal{M}$) to equivalent errors within $\mathcal{M}$ using a coupling argument. This leads us to introduce the CVaR-VI-DISC algorithm, which is tailored for practical applications.

The following theorem guarantees that the optimization error assumption is met when when Algorithm 4 is employed.

Theorem 4.9. *By selecting $\eta \leq \epsilon\tau/3H$, we ensure that*

$$|CVaR_\tau^{\rho^\star}(s_1; r) - CVaR_\tau^{\hat{\rho}}(s_1; r)| \leq \epsilon/3, \tag{10}$$

where $\rho^\star$ represents the policy generated by Algorithm 3 and $\hat{\rho}$ is the output of Algorithm 4. Consequently, the optimization error is bounded by $\epsilon/3$, which satisfies Assumption 4.1.

Proof. Please refer to the Appendix for more details. □

4.3. Adaptability to Varying Risk Tolerances

We further introduce an important proposition that underscores the adaptability of our exploration process to different levels of risk tolerance τ:

Proposition 4.10. *For any $\tau' \geq \tau$, the exploration dataset obtained through Algorithm 1 at risk tolerance τ contains the requisite information for conducting CVaR-RF RL with any higher risk tolerance τ'. Consequently, the planning phase is also compatible with any given $\tau' \geq \tau$.*

Proof. Utilizing Lemma 4.2, we observe that as $\epsilon\tau/3 \leq \epsilon\tau'/3$, the CVaR-RF exploration algorithm configured with a risk tolerance of τ also satisfies the (ϵ, δ)-PAC criterion for CVaR-RF RL when operating under a higher risk tolerance $\tau' \geq \tau$. Furthermore, invoking Theorem 4.9, we have that the stipulated optimization error condition is met since $\eta \leq \eta'$. This implies that the planning phase remains efficacious under these adjusted parameters. □

5. Lower Bound

In this section, we develop a lower bound of the sample complexity for CVaR-RF exploration. We present a theorem that delineates this lower bound, applicable to any algorithm operating within the CVaR-RF exploration framework.

Theorem 5.1. *Consider a universal constant $C > 0$. For a given risk tolerance $\tau \in (0, 1]$, if the number of actions $A \geq 2$, the number of states $S \geq C \log_2 A + 2$, the horizon $H \geq C \log_2 S + 1$, and the accuracy parameter $\epsilon \leq \min\{1/4\tau, H/48\tau\}$, then any CVaR-RF exploration algorithm that can output ϵ-optimal policies for an arbitrary number of adaptively chosen reward functions*

with a success probability $\delta = 1/2$ must collect at least $\Omega(S^2AH^2/\tau\epsilon^2)$ trajectories in expectation.

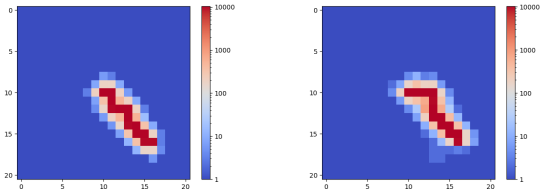
Proof Sketch. Here we highlight the main idea of our lower bound proof, while the detailed proof can be found in the Appendix. Our proof is inspired by the lower bound construction in for the reward-free RL (Jin et al., 2020). The key idea is that any reward-free risk neutral problem can be transformed into a CVaR-RF RL problem. If a CVaR-RF exploration algorithm that can output ϵ-optimal policies in the transformed CVaR-RF RL problem, it can also solve the original reward-free risk neutral problem. Specifically, for a MDP $\mathcal{M}$ with initial state s_1, we consider a new MDP $\mathcal{M}'$ with an initial state s_0. For any action a, $\mathbb{P}(s_1|s_0,a) = \tau$, $\mathbb{P}(s'|s_0,a) = 1-\tau$, $\mathbb{P}(s'|s',a) = 1$, and $r(s',a) = 1$. For any adaptively chosen reward function for $\mathcal{M}$ and a policy π, the CVaR with tolerance τ following policy in the new MDP $\mathcal{M}'$ is equal to the cumulative rewards in the original MDP $\mathcal{M}$. (Jin et al., 2020) shows that any reward-free exploration algorithm that output ϵ-optimal policy from initial state s_1 must collect at least $\Omega(S^2AH^2/\tau\epsilon^2)$ trajectories in expectation. Thus, from the initial state s_0, the CVaR-RF exploration algorithm must collect at least $\Omega(S^2AH^2/\tau\epsilon^2)$ trajectories in expectation. □

This theorem illustrates that, compared with the lower bound, the upper bound established in Theorem 4.7 has by an additional factor of H^2 and $1/\tau$, while being tight with respect to the parameters S, A, ϵ. If τ is a constant, our result is nearly minimax-optimal with an additional factor on H^2. An interesting direction of the future work is utilizing the empirical Bernstein inequality to further improve the sample complexity. The H factor can potentially be optimized by adopting an approach similar to (Ménard et al., 2021) by introducing an empirical Bernstein inequality derived from a control of the transition probability. As shown in (Wang et al., 2023), the Bernstein inequality could also potentially improve the dependence on τ under a continuity assumption. Furthermore, compared with the risk-neutral reward-free RL, our derived lower bound for any CVaR-RF exploration algorithm includes an additional τ in the denominator. This is because CVaR focuses on the τ worst outcomes. Additionally, the CVaR setting poses challenges due to non-Markovianity, requiring more efforts in achieving a minimax optimal sample complexity bound.

6. Experiments

In this section, we provide numerical examples to evaluate the proposed CVaR-RF RL framework. In these examples, we use similar experimental setup as in (Kaufmann et al., 2021). Our environment is configured as a grid-world consisting of 21×21 states, where each state offers four possible actions (up, down, left, right), and actions leading to the boundary result in remaining in the current state. The agent will move to the correct state with a probability of 0.95. However, there is an equal probability of $\frac{0.05}{3}$ for the agent to move in any one of the other three directions. Initially, the exploration algorithm CVaR-RF-UCRL runs without reward information, collecting $n = 30,000$ transitions. The empirical transition probability $\hat{\mathbb{P}}$ is then estimated. We use the $\beta(n,\delta)$ threshold from Theorem 4.7 with $\delta = 0.1$ and set a time horizon H of 20. Using the obtained dataset and $\hat{\mathbb{P}}$, the planning algorithm derives near-optimal policies, employing CVaR-VI-DISC as the solver.

Reward Setup 1: The first one is similar with (Kaufmann et al., 2021), where the agent starts at position $(10,10)$. The reward structure is primarily set at 0 for most states, except at $(16,16)$ where it is 1.0. Here we choose $\epsilon = 0.1$. Then we executing the output policy of CVaR-VI-DISC in the same grid-world for $K = 10,000$ trajectories and plot the number of state visits following the policy. For comparison, we also generate the optimal policy using true transition probability. Figures 1a displays the number of visits to each state following the policy generated from $\mathbb{P}$, while Figure 1b shows for $\hat{\mathbb{P}}$. Additionally, Table 1 presents the CVaR values under both true and empirical transition probabilities.



(a) Optimal policy (b) CVaR-VI-DISC

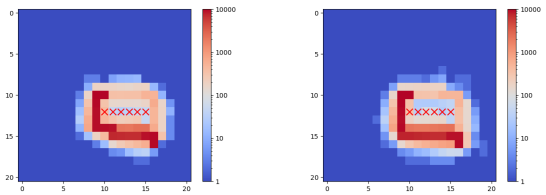
Figure 1. Number of state visits following policies generated under $\mathbb{P}$ and $\hat{\mathbb{P}}$ in reward setup 1 with risk tolerance $\tau = 0.05$.

ϵ, τ	$\text{CVaR}_{\mathbb{P}}$	$\text{CVaR}_{\hat{\mathbb{P}}}$	Error
$0.1, 0.05$	4.308	4.258	0.05
$0.1, 0.95$	4.960	4.954	0.006

Table 1. CVaR values under reward setup 1 with different τ.

These visitation patterns, shown in Figures 1a and 1b, are notably similar, indicating that the agent tends to favor states with higher rewards. This behavior is consistent with the objective of maximizing CVaR. The similarity in patterns under both true and empirical transition probabilities underscores the reliability of the data collected during the exploration phase. Moreover, with $\epsilon = 0.1$ and $\tau = 0.05$, the difference between true and empirical CVaR is -0.05, which is below the anticipated error threshold of $\epsilon = 0.1$. Similarly, with $\epsilon = 0.1$ and $\tau = 0.95$, the error is only 0.006, again less than the threshold of 0.1. These results align with our theoretical analysis.

Reward Setup 2: We consider a more complex case as the reward structure is primarily set at 0.5 for most states, except at $(16, 16)$ where it is 1.0, and a zero-reward zone marked 'x' from $(12, 10)$ to $(12, 16)$.



(a) Optimal policy (b) CVaR-VI-DISC

Figure 2. Number of state visits following policies generated under $\mathbb{P}$ and $\hat{\mathbb{P}}$ in reward setup 2 with risk tolerance $\tau = 0.05$.

ϵ, τ	$\text{CVaR}_{\mathbb{P}}$	$\text{CVaR}_{\hat{\mathbb{P}}}$	Error
$0.1, 0.05$	1.852	1.829	0.023
$0.1, 0.95$	1.993	1.990	0.003

Table 2. CVaR values under reward setup 2 with different τ.

Figure 2 and Table 2 illustrate that CVaR-RF RL effectively avoids traversing zero-reward regions, and the observed errors remain within the pre-defined thresholds. These outcomes are also consistent with the CVaR's property as the agent is more risk-averse compared to risk-neutral case.

7. Conclusion

In this paper, we have introduced a novel risk-sensitive reward-free RL framework based on CVaR (CVaR-RF RL), which is able to solve CVaR RL for given any reward function after a singular reward-free exploration phase. We have proposed CVaR-RF-UCRL as the exploration algorithm and established upper and lower bounds for the sample complexity. We have developed a CVaR-RF-planning algorithm, equipped with CVaR-VI and CVaR-VI-DISC to generate near-optimal Markov policies solely based on the exploration dataset and given reward function. Through our numerical experiments, we have validated the effectiveness and practicality of this CVaR-RF-RF framework.

Acknowledgement

This work was supported by the National Science Foundation under Grants CCF-21-12504 and CCF-22-32907.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Acerbi, C. and Tasche, D. On the Coherence of Expected Shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. Coherent Measures of Risk. *Mathematical Finance*, 9(3):203–228, 1999.

Auer, P., Jaksch, T., and Ortner, R. Near-Optimal Regret Bounds for Reinforcement Learning. *Advances in Neural Information Processing Systems*, 21, 2008.

Azar, M. G., Osband, I., and Munos, R. Minimax Regret Bounds for Reinforcement Learning. In *Proc. International Conference on Machine Learning*, pp. 263–272, Sydney, Australia, 2017. PMLR.

Bastani, O., Ma, J. Y., Shen, E., and Xu, W. Regret Bounds for Risk-Sensitive Reinforcement Learning. *Advances in Neural Information Processing Systems*, 35:36259–36269, 2022.

Bäuerle, N. and Ott, J. Markov Decision Processes with Average-Value-at-Risk Criteria. *Mathematical Methods of Operations Research*, 74:361–379, 2011.

Bellemare, M. G., Dabney, W., and Rowland, M. *Distributional Reinforcement Learning*. MIT Press, 2023.

Chen, J., Modi, A., Krishnamurthy, A., Jiang, N., and Agarwal, A. On the Statistical Efficiency of Reward-Free Exploration in Non-Linear RL. *Advances in Neural Information Processing Systems*, 35:20960–20973, 2022.

Chow, Y. and Ghavamzadeh, M. Algorithms for CVaR Optimization in MDPs. *Advances in Neural Information Processing Systems*, 27, 2014.

Chow, Y., Tamar, A., Mannor, S., and Pavone, M. Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach. *Advances in Neural Information Processing Systems*, 28, 2015.

Chow, Y., Ghavamzadeh, M., Janson, L., and Pavone, M. Risk-Constrained Reinforcement Learning with Percentile Risk Criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

Dann, C. and Brunskill, E. Sample Complexity of Episodic Fixed-Horizon Reinforcement Learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Dann, C., Lattimore, T., and Brunskill, E. Unifying PAC andRregret: Uniform PAC bounds for Episodic Reinforcement Learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Delage, E. and Mannor, S. Percentile Optimization for Markov Decision Processes with Parameter Uncertainty. *Operations Research*, 58(1):203–213, 2010.

Ding, Y., Jin, M., and Lavaei, J. Non-Stationary Risk-Sensitive Reinforcement Learning: Near-Optimal Dynamic Regret, Adaptive Detection, and Separation Design. In *Proc. the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7405–7413, Washington, DC, 2023.

Du, Y., Wang, S., and Huang, L. Provably Efficient Risk-Sensitive Reinforcement Learning: Iterated CVaR and Worst Path. In *Proc. The Eleventh International Conference on Learning Representations*, 2022.

Fei, Y., Yang, Z., Chen, Y., Wang, Z., and Xie, Q. Risk-Sensitive Reinforcement Learning: Near-Optimal Risk-Sample Tradeoff in Regret. *Advances in Neural Information Processing Systems*, 33:22384–22395, 2020.

Fei, Y., Yang, Z., and Wang, Z. Risk-Sensitive Reinforcement Learning with Function Approximation: A Debiasing Approach. In *Proc. International Conference on Machine Learning*, pp. 3198–3207. PMLR, 2021.

Fiechter, C.-N. Efficient Reinforcement Learning. In *Proc. The Seventh Annual Conference on Computational Learning Theory*, pp. 88–97, New Brunswick, NJ, 1994.

Hau, J. L., Petrik, M., and Ghavamzadeh, M. Entropic Risk Optimization in Discounted MDPs. In *Proc. International Conference on Artificial Intelligence and Statistics*, pp. 47–76, Valencia, Spain, 2023. PMLR.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-Learning Provably Efficient? *Advances in Neural Information Processing Systems*, 31, 2018.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-Free Exploration for Reinforcement Learning. In *Proc. International Conference on Machine Learning*, pp. 4870–4879, Stockholm, Sweden, 2020. PMLR.

Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. Adaptive Reward-Free Exploration. In *Proc. Algorithmic Learning Theory*, pp. 865–891, Paris, France, 2021. PMLR.

Keramati, R., Dann, C., Tamkin, A., and Brunskill, E. Being Optimistic to Be Conservative: Quickly Learning a CVaR Policy. In *Proc. the AAAI conference on artificial intelligence*, volume 34, pp. 4436–4443, New York, NY, 2020.

La, P. and Ghavamzadeh, M. Actor-Critic Algorithms for Risk-Sensitive MDPs. *Advances in Neural Information Processing Systems*, 26, 2013.

Li, L. Sample Somplexity Bounds of Exploration. In *Reinforcement Learning: State-of-the-Art*, pp. 175–204. Springer, 2012.

Ménard, P., Domingues, O. D., Jonsson, A., Kaufmann, E., Leurent, E., and Valko, M. Fast Active Learning for Pure Exploration in Reinforcement Learning. In *Proc. International Conference on Machine Learning*, pp. 7599–7608. PMLR, 2021.

Miryoosefi, S. and Jin, C. A Simple Reward-Free Approach to Constrained Reinforcement Learning. In *Proc. International Conference on Machine Learning*, pp. 15666–15698, Baltimore, MD, 2022. PMLR.

Ni, X. and Lai, L. Policy Gradient Based Entropic-VaR Optimization in Risk-Sensitive Reinforcement Learning. In *Proc. 2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1–6, Allerton, IL, 2022a. IEEE.

Ni, X. and Lai, L. Risk-Sensitive Reinforcement Learning via Entropic-VaR Optimization. In *Proc. 2022 56th Asilomar Conference on Signals, Systems, and Computers*, pp. 953–959, Pacific Grove, CA, 2022b. IEEE.

Prashanth, L., Fu, M. C., et al. Risk-Sensitive Reinforcement Learning via Policy Gradient Search. *Foundations and Trends® in Machine Learning*, 15(5):537–693, 2022.

Rockafellar, R. T., Uryasev, S., et al. Optimization of Conditional Value-at-Risk. *Journal of Risk*, 2:21–42, 2000.

Shen, Y., Tobia, M. J., Sommer, T., and Obermayer, K. Risk-Sensitive Reinforcement Learning. *Neural Computation*, 26(7):1298–1328, 2014.

Tamar, A., Chow, Y., Ghavamzadeh, M., and Mannor, S. Policy Pradient for Coherent Risk Measures. *Advances in Neural Information Processing Systems*, 28, 2015a.

Tamar, A., Glassner, Y., and Mannor, S. Optimizing the CVaR via Sampling. In *Proc. the AAAI Conference on Artificial Intelligence*, volume 29, Austin, TX, 2015b.

Wang, K., Kallus, N., and Sun, W. Near-minimax-optimal risk-sensitive reinforcement learning with cvar. *arXiv preprint arXiv:2302.03201*, 2023.

Wang, R., Du, S. S., Yang, L., and Salakhutdinov, R. R. On Reward-Free Reinforcement Learning with Linear Function Approximation. *Advances in Neural Information Processing Systems*, 33:17816–17826, 2020.

Ying, C., Zhou, X., Su, H., Yan, D., Chen, N., and Zhu, J. Towards Safe Reinforcement Learning via Constraining Conditional Value-at-Risk. *arXiv preprint arXiv:2206.04436*, 2022.

Zhang, Z., Du, S., and Ji, X. Near Optimal Reward-Free Reinforcement Learning. In *Proc. International Conference on Machine Learning*, pp. 12402–12412. PMLR, 2021.

A. Proof of Exploration Phase

A.1. Proof of Lemma 4.2

Recall the definition of value function V for various policy types:

$$\pi \in \Pi_{\mathcal{H}} : V_h^{\pi}(s_h, b_h; \mathcal{H}_h) = \mathbb{E}_{\pi}\left[\left(b_h - \sum_{t=h}^{H} r_t\right)^+ \middle| s_h, b_h, \mathcal{H}_h\right],$$

$$\rho \in \Pi^{\text{Aug}} : V_h^{\rho}(s_h, b_h) = \mathbb{E}^{\rho}\left[\left(b_h - \sum_{t=h}^{H} r_t\right)^+ \middle| s_h, b_h\right].$$

Notice that executing ρ, b in the augmented MDP is equivalent to executing policy $\pi^{\rho,b}$ in the original MDP, where $\pi_h^{\rho,b}(s_h, \mathcal{H}_h) = \rho_h(s_h, b - r_1 - \ldots - r_{h-1})$. Consequently, their V functions should be equivalent.

Therefore, by Lemma D.3, we have

$$\begin{aligned}
&\text{CVaR}_\tau^\star(s_1;r) - \text{CVaR}_\tau^{\hat{\rho}_r}(s_1;r) \\
&= \text{CVaR}_\tau^{\pi^{\rho^\star,b_1^\star}}(s_1;r) - \text{CVaR}_\tau^{\pi^{\hat{\rho},\hat{b}_1}}(s_1;r) \\
&= \underbrace{\text{CVaR}_\tau^{\pi^{\rho^\star,b_1^\star}}(s_1;r) - \widehat{\text{CVaR}}_\tau^{\pi^{\rho^\star,b_1^\star}}(s_1;r)}_{\text{Evaluation error I}} + \underbrace{\widehat{\text{CVaR}}_\tau^{\pi^{\rho^\star,b_1^\star}}(s_1;r) - \widehat{\text{CVaR}}_\tau^{\pi^{\hat{\rho}^\star,\hat{b}_1^\star}}(s_1;r)}_{\leq 0 \text{ by definition}} \\
&\quad + \underbrace{\widehat{\text{CVaR}}_\tau^{\pi^{\hat{\rho}^\star,\hat{b}_1^\star}}(s_1;r) - \widehat{\text{CVaR}}_\tau^{\pi^{\hat{\rho},\hat{b}_1}}(s_1;r)}_{\text{optimization error } \leq \epsilon/3 \text{ by Assumption 4.1}} + \underbrace{\widehat{\text{CVaR}}_\tau^{\pi^{\hat{\rho},\hat{b}_1}}(s_1;r) - \text{CVaR}_\tau^{\pi^{\hat{\rho},\hat{b}_1}}(s_1;r)}_{\text{Evaluation error II}}.
\end{aligned}$$

By the triangle inequality, we have

$$\begin{aligned}
&\left|\text{CVaR}_\tau^\star(s_1;r) - \text{CVaR}_\tau^{\hat{\rho}_r^\star}(s_1;r)\right| \\
&\leq \left|\text{CVaR}_\tau^{\pi^{\rho^\star,b^\star}}(s_1;r) - \widehat{\text{CVaR}}_\tau^{\pi^{\rho^\star,b^\star}}(s_1;r)\right| + \left|\widehat{\text{CVaR}}_\tau^{\pi^{\hat{\rho}^\star,\hat{b}^\star}}(s_1;r) - \text{CVaR}_\tau^{\pi^{\hat{\rho}^\star,\hat{b}^\star}}(s_1;r)\right|.
\end{aligned}$$

For the evaluation errors, by the definition of CVaR, we have

$$\begin{aligned}
\left|\text{CVaR}_\tau^{\pi^{\rho^\star,b_1^\star}}(s_1;r) - \widehat{\text{CVaR}}_\tau^{\pi^{\rho^\star,b_1^\star}}(s_1;r)\right| &= \left|b_1^\star - \tau^{-1}V_1^{\pi^{\rho^\star,b_1^\star}}(s_1,b_1^\star;r) - \max_{b_1\in[0,H]}\left\{b_1 - \tau^{-1}\hat{V}_1^{\pi^{\rho^\star,b_1^\star}}(s_1,b_1;r)\right\}\right| \\
&\leq \left|b_1^\star - \tau^{-1}V_1^{\pi^{\rho^\star,b_1^\star}}(s_1,b_1^\star;r) - \left(b_1^\star - \tau^{-1}\hat{V}_1^{\pi^{\rho^\star,b_1^\star}}(s_1,b_1^\star;r)\right)\right| \\
&\leq \tau^{-1}\left|V_1^{\pi^{\rho^\star,b_1^\star}}(s_1,b_1^\star;r) - \hat{V}_1^{\pi^{\rho^\star,b_1^\star}}(s_1,b_1^\star;r)\right|,
\end{aligned}$$

and similarly,

$$\left|\widehat{\text{CVaR}}_\tau^{\pi^{\hat{\rho},\hat{b}_1}}(s_1;r) - \text{CVaR}_\tau^{\pi^{\hat{\rho},\hat{b}_1}}(s_1;r)\right| \leq \tau^{-1}\left|V_1^{\pi^{\hat{\rho},\hat{b}_1}}(s_1,\hat{b}_1;r) - \hat{V}_1^{\pi^{\hat{\rho},\hat{b}_1}}(s_1,\hat{b}_1;r)\right|.$$

Therefore, if an exploration algorithm that satisfies

$$\left|V_1^{\rho}(s_1,b_1;r) - \hat{V}_1^{\rho}(s_1,b_1;r)\right| \leq \epsilon\tau/3, \forall\rho \in \Pi^{\text{Aug}}, \forall b_1 \in [0,H],$$

or equivalently,

$$\left|Q_1^{\rho}(s_1,b_1,\rho(s_1,b_1);r) - \hat{Q}_1^{\rho}(s_1,b_1,\rho(s_1,b_1);r)\right| \leq \epsilon\tau/3, \forall\rho \in \Pi^{\text{Aug}}, \forall b_1 \in [0,H],$$

it further ensures $\left|\text{CVaR}_\tau^\star(s_1;r) - \text{CVaR}_\tau^{\hat{\rho}_r^\star}(s_1;r)\right| \leq \epsilon$, which completes the proof.

A.2. Proof of Lemma 4.5

We first consider the case where the initial budget b_1 is fixed and for convenience, we omit the index $h+1$ by using (s', b'). Referring to the Bellman equations in both the empirical augmented MDP and the true augmented MDP,

$$\hat{Q}_h^{t,\rho}(s_h, b_h, a_h; r) = \sum_{s'} \hat{\mathbb{P}}_h^t(s'|s,a)\hat{Q}_{h+1}^{t,\rho}(s', b', \rho(s', b'); r),$$

$$\text{and} \quad Q_h^{\rho}(s_h, b_h, a_h; r) = \sum_{s'} \mathbb{P}_h(s'|s,a)Q_{h+1}^{\rho}(s', b', \rho(s', b'); r),$$

we have

$$\begin{aligned}
&\hat{Q}_h^{t,\rho}(s_h, b_h, a_h; r) - Q_h^{\rho}(s_h, b_h, a_h; r) \\
&= \sum_{s'} \hat{\mathbb{P}}_h^t(s'|s,a)\hat{Q}_{h+1}^{t,\rho}(s', b', \rho(s', b'); r) - \sum_{s'} \mathbb{P}_h(s'|s,a)Q_{h+1}^{\rho}(s', b', \rho(s', b'); r) \\
&= \sum_{s'} \left(\hat{\mathbb{P}}_h^t(s'|s,a) - \mathbb{P}_h(s'|s,a)\right) Q_{h+1}^{\rho}(s', b', \rho(s', b'); r) \\
&\quad + \sum_{s'} \hat{\mathbb{P}}_h^t(s'|s,a) \left(\hat{Q}_{h+1}^{t,\rho}(s', b', \rho(s', b'); r) - Q_{h+1}^{\rho}(s', b', \rho(s', b'); r)\right).
\end{aligned}$$

Thus, for $n_h^t(s,a) \geq 0$, we obtain

$$\begin{aligned}
&\hat{e}_h^{t,\rho}(s_h, b_h, a_h; r) \\
&= |\hat{Q}_h^{t,\rho}(s_h, b_h, a_h; r) - Q_h^{\rho}(s_h, b_h, a_h; r)| \\
&\overset{(1)}{\leq} \sum_{s'} \left|\hat{\mathbb{P}}_h^t(s'|s,a) - \mathbb{P}_h(s'|s,a)\right| Q_{h+1}^{\rho}(s', b', \rho(s', b'); r) \\
&\quad + \sum_{s'} \hat{\mathbb{P}}_h^t(s'|s,a) \left|\hat{Q}_{h+1}^{t,\rho}(s', b', \rho(s', b'); r) - Q_{h+1}^{\rho}(s', b', \rho(s', b'); r)\right| \\
&\overset{(2)}{\leq} b_1 \|\hat{\mathbb{P}}_h^t(\cdot|s,a) - \mathbb{P}_h(\cdot|s,a)\|_1 + \sum_{s'} \hat{\mathbb{P}}_h^t(s'|s,a)\hat{e}_{h+1}^{t,\rho}(s', b', a'; r) \\
&\overset{(3)}{\leq} b_1 \sqrt{\frac{2\beta(n_h^t(s,a), \delta)}{n_h^t(s,a)}} + \sum_{s'} \hat{\mathbb{P}}_h^t(s'|s,a)\hat{e}_{h+1}^{t,\rho}(s', b', a'; r),
\end{aligned}$$

where (1) is due to the Pinsker's inequality; (2) is due to the fact that $Q_h^{\rho}(s_h, b_h, a_h; r) \leq b_1$ ($Q_h^{\rho}(s_h, b_h, a_h; r) \leq (b_h)^+ \leq b_1$ as $b_{h+1} = b_h - r_h$) for all s, a, b, r and the definition of L_1 norm; (3) is due to the fact that $\text{TV}(P, Q) = \frac{1}{2}\|P(\cdot) - Q(\cdot)\|_1 \leq \sqrt{\frac{1}{2}\text{KL}(P, Q)}$ and the definition of $\mathcal{E}$.

Notice that $\hat{e}_h^{t,\rho}(s_h, b_h, a_h; r) \leq b_1$, then for all $n_h^t(s,a) \geq 0$, we have

$$\hat{e}_h^{t,\rho}(s_h, a_h, b_h; r) \leq \min\left\{b_1, b_1 \sqrt{\frac{2\beta(n_h^t(s,a), \delta)}{n_h^t(s,a)}} + \sum_{s'} \hat{\mathbb{P}}_h^t(s'|s,a)\hat{e}_{h+1}^{t,\rho}(s', a', b'; r)\right\}.$$

Notice that $b_1 \in [0, H]$, in order to find the upper bound of the estimation error over all the initial budgets, we extend the inequality to

$$\begin{aligned}
\hat{e}_h^{t,\rho}(s_h, a_h, b_h; r) &\leq \max_{b_1 \in [0,H]} \left\{\min\left\{b_1, b_1 \sqrt{\frac{2\beta(n_h^t(s,a), \delta)}{n_h^t(s,a)}} + \sum_{s'} \hat{\mathbb{P}}_h^t(s'|s,a)\hat{e}_{h+1}^{t,\rho}(s', a', b'; r)\right\}\right\} \\
&\leq \min\left\{H, H \sqrt{\frac{2\beta(n_h^t(s,a), \delta)}{n_h^t(s,a)}} + \sum_{s'} \hat{\mathbb{P}}_h^t(s'|s,a)\hat{e}_{h+1}^{t,\rho}(s', a', b'; r)\right\}.
\end{aligned}$$

Now we prove Lemma 4.5 by induction. For $H+1$, since

$$\hat{e}^{t,\rho}_{H+1}(s,a,b;r) = |\hat{Q}^{t,\rho}_{H+1}(s,a,b;r) - Q^{\rho}_{H+1}(s,a,b;r)| = \max\{0,b_1\} - \max\{0,b_1\} = 0$$

and $E^t_{H+1}(s,a) = 0$ for all (s,a), the result is true. Assume the result holds for $h+1$, i.e., $\hat{e}^{t,\rho}_{h+1}(s,a,b;r) \le E^t_{h+1}(s,a;b_1)$ for all (s,a), we have

$$\begin{aligned}\hat{e}^{t,\rho}_h(s,a,b;r) &\le \min\left\{H, H\sqrt{\frac{2\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)}} + \sum_{s'}\hat{\mathbb{P}}^t_h(s'|s,a)\hat{e}^{t,\rho}_{h+1}(s',a',b';r)\right\}\\ &\le \min\left\{H, H\sqrt{\frac{2\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)}} + \sum_{s'}\hat{\mathbb{P}}^t_h(s'|s,a)\max_{a\in\mathcal{A}}E_{h+1}(s',a)\right\} = E^t_h(s,a)\end{aligned}$$

holds for h, which complete the proof.

A.3. Proof of Theorem 4.7

Notice that in the exploration phase, we follow the exploration policy π rather than ρ. We begin by introducing some notations. Let $\mathbb{P}^{\pi}_h(s,a)$ represent the probability that the state-action pair (s,a) is reached at the h-th step of a trajectory under the exploraion policy π. We use the shorthand $p^h_t(s,a) = p^h_{\pi_t}(s,a)$ for simplicity. The pseudo-counts $\bar{n}^t_h(s,a)$ are defined as $\sum_{i=1}^t \mathbb{P}^i_h(s,a)$, and we define the event

$$\mathcal{E}^{\text{cnt}} = \left\{\forall t\in\mathbb{N}^\star, \forall h\in[H], \forall(s,a)\in\mathcal{S}\times\mathcal{A} : n^t_h(s,a) \ge \frac{1}{2}\bar{n}^t_h(s,a) - \beta^{\text{cnt}}(\delta)\right\},$$

where $\beta_{\text{cnt}}(\delta) = \log(2SAH/\delta)$. Recalling the event $\mathcal{E}$ defined in Lemma 4.5, we let $\mathcal{F} = \mathcal{E}\cap\mathcal{E}^{\text{cnt}}$ and introduce the following lemma.

By Lemma D.4 and the principle of inclusion-exclusion, we have $\mathbb{P}(\mathcal{F}) = \mathbb{P}(\mathcal{E}\cap\mathcal{E}^{\text{cnt}}) = \mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{E}^{\text{cnt}}) - \mathbb{P}(\mathcal{E}\cup\mathcal{E}^{cnt}) \ge \mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{E}^{\text{cnt}}) - 1 = 1-\delta$. From Lemma 4.6, on the event $\mathcal{F}$, it is shown that $\text{CVaR}^{\star}_{\tau}(s_1;r) - \text{CVaR}^{\hat{\rho}^*}_{\tau}(s_1;r) \le \epsilon$ for all reward functions r, thereby proving that CVaR-RF-UCRL is (ε,δ)-PAC.

We now proceed to upper bound the sample complexity of CVaR-RF-UCRL on the event $\mathcal{F}$. The first step involves introducing an average upper bound on the error at step h under policy π^{t+1}, defined as

$$\mathbb{Q}^t_h = \sum_{(s,a)}\mathbb{P}^{t+1}_h(s,a)E^t_h(s,a).$$

By Lemma D.1, the average errors can be related as follows:

$$\begin{aligned}\mathbb{Q}^h_t &\le 3H\sum_{(s,a)}\mathbb{P}^{t+1}_h(s,a)\left[\sqrt{\frac{\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)}}\wedge 1\right] + \sum_{(s,a)}\sum_{(s',a')}\mathbb{P}^{t+1}_h(s,a)\mathbb{P}_h(s'|s,a)\mathbb{I}(a'=\pi^{t+1}(s'))E^t_{h+1}(s',a')\\ &\le 3H\sum_{(s,a)}\mathbb{P}^{t+1}_h(s,a)\left[\sqrt{\frac{\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)}}\wedge 1\right] + \mathbb{Q}^t_{h+1}.\end{aligned}$$

For $h=1$, observe that $\mathbb{P}^{t+1}_1(s_1,a)E^t_1(s_1,a) = E^t_1(s_1,\pi^{t+1}_1(s_1))\mathbb{I}(\pi^{t+1}_1(s_1)=a)$, as the policy is deterministic. Now, if $t < t_{\text{stop}}$, $E^t_1(s_1,\pi^{t+1}_1(s_1)) \ge \epsilon/3$ by definition of the stopping rule, hence $\mathbb{Q}^1_t = \sum_a \mathbb{P}^{t+1}_1(s_1,a)E^t_1(s_1,a) \ge (\epsilon\tau/3)\sum_{a\in\mathcal{A}}\mathbb{I}(\pi^{t+1}_1(s_1)=a) = \epsilon\tau/3$. Thus, we have

$$\frac{\epsilon\tau}{3} \le 3\sum_{h=1}^H\sum_{(s,a)}H\mathbb{P}^{t+1}_h(s,a)\left[\sqrt{\frac{\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)}}\wedge 1\right]$$

for $t < t_{\text{stop}}$. Summing these inequalities for $t\in\{0,\dots,T\}$ where $T < t_{\text{stop}}$ gives:

$$(T+1)\epsilon\tau \le 9\sum_{h=1}^H H\sum_{(s,a)}\sum_{t=0}^T\mathbb{P}^{t+1}_h(s,a)\left[\sqrt{\frac{\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)}}\wedge 1\right].$$

The next step involves relating the counts to the pseudo-counts, taking into account that the event $\mathcal{E}^{\text{cnt}}$ holds.

Using Lemma D.5, it can be stated that, on the event F, for $T < t_{\text{stop}}$, the inequality

$$\begin{aligned}(T+1)\epsilon\tau &\leq 18\sum_{h=1}^{H} H \sum_{(s,a)} \sum_{t=0}^{T} \mathbb{P}_h^{t+1}(s,a)\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)\vee 1}} \\ &\leq 18\sqrt{\beta(T+1,\delta)}\sum_{h=1}^{H} H \sum_{(s,a)} \sum_{t=0}^{T} \frac{\bar{n}_h^{t+1}(s,a) - n_h^t(s,a)}{\sqrt{\bar{n}_h^t(s,a)\vee 1}},\end{aligned}$$

is derived, where the relation $\mathbb{P}_h^{t+1}(s,a) = \bar{n}_h^{t+1}(s,a) - \bar{n}_h^t(s,a)$, as per the definition of pseudo-counts, is used.

Applying Lemma D.6 to bound the sum over t, we get:

$$\begin{aligned}(T+1)\epsilon\tau &\leq 18(1+\sqrt{2})\sqrt{\beta(T+1,\delta)}\sum_{h=1}^{H} H \sum_{(s,a)} \sqrt{n_h^{T+1}(s,a)} \\ &\leq 18(1+\sqrt{2})\sqrt{\beta(T+1,\delta)}\sum_{h=1}^{H} H\sqrt{SA}\sqrt{\sum_{s,a} n_h^{T+1}(s,a)}.\end{aligned}$$

Given that $\sum_{s,a} n_h^{T+1}(s,a) = T+1$, the inequality simplifies to:

$$\sqrt{T+1}\epsilon\tau \leq 18(1+\sqrt{2})\sqrt{SA}H^2\sqrt{\beta(T+1,\delta)}.$$

For sufficiently large T, this inequality cannot hold, as the left-hand side grows with $\sqrt{T}$, while the right-hand side is logarithmic. Therefore, t_{stop} is finite and satisfies (applying the inequality to $T = t_{\text{stop}} - 1$):

$$t_{\text{stop}} \leq \tilde{\mathcal{O}}\left(\frac{H^4S^2A}{\epsilon^2\tau^2}\right)$$

The conclusion follows from Lemma D.7.

B. Proof of Planning Phase

B.1. Proof of Theorem 4.8

The utilization of discretization in the algorithm significantly impacts its computational tractability, and it is applied in two main areas:
1. In the dynamic programming step at each timestep h, the algorithm exclusively computes $Q_h(s_h, b_h, a_h)$ for all s_h, a_h and b_h within the grid. This leads to a total runtime of $\mathcal{O}(SAH\eta^{-1}T_{\text{step}})$, where T_{step} represents the time required for each step. The time complexity here arises from discretization and is a function of the state space size, action space size, and the horizon length.
2. When computing $\hat{b}$, the algorithm searches over the grid to find the solution. Since the returns distribution is supported on the grid, the τ-quantile of the return distribution (the optimal solution) exists on the grid. This computation has a time complexity of $\mathcal{O}(\eta^{-1})$, which is considered a lower-order term compared to the first part.

It's important to note that the most time-consuming part of the algorithm is the computation of expectations, specifically the term:

$$[\mathbb{P}_h V_{h+1}](s_h, b_h, a_h) = \mathbb{E}_{s_{h+1}\sim\mathbb{P}(\cdot|s_h,a_h)}[V_{h+1}^{\star}(s_{h+1}, b_{h+1})].$$

In the discretized MDP, this expectation can be computed using only grid elements, implying $T_{\text{step}} = \mathcal{O}(S\eta^{-1})$. As a result, the overall time complexity of this algorithm is approximately $\mathcal{O}(SAH\eta^{-1}T_{\text{step}}) = \mathcal{O}(S^2AH\eta^{-2})$.

B.2. Proof of Theorem 4.9

The proof draws inspiration from (Bastani et al., 2022; Wang et al., 2023). To facilitate the discussion, we introduce the following notation. Let $Z_{\rho,\mathcal{M}}$ represent the returns from executing ρ in the MDP $\mathcal{M}$. For random variables X, Y,

we say Y stochastically dominates X, which is denoted $X \preceq Y$. This dominance implies that for any real value t, the probability that Y is less than or equal to t is greater than or equal to the probability of X being less than or equal to t, i.e., $\forall t \in \mathbb{R} : Pr(Y \leq t) \leq Pr(X \leq t)$.

1) From disc($\mathcal{M}$) to $\mathcal{M}$:

Consider any policy $\rho \in \Pi^{\text{Aug}}$ and $b \in [0, 1]$ (which we use in disc($\mathcal{M}$)). Define an adapted policy for use in $\mathcal{M}$ as follows:

$$\text{adapted}(\rho, b_1)_h(s_h, r_{1:h-1}) = \rho_h(s_h, b_1 - \phi(r_1) - \cdots - \phi(r_h - 1)).$$

The adapted policy simulates the evolution of b in disc($\mathcal{M}$) by using the history. Let $Z_{\rho,b,\text{disc}(\mathcal{M})}$ be the returns from running ρ, b in disc($\mathcal{M}$). Let $Z_{\text{adopted}(\rho,b),\mathcal{M}}$ be the returns from running adopted(ρ, b) in $\mathcal{M}$. According to Lemma H.1 in (Wang et al., 2023), we almost surely have

$$Z_{\rho,b,\text{disc}(\mathcal{M})} - H\eta \preceq Z_{\text{adopted}(\rho,b),\mathcal{M}} \preceq Z_{\rho,b,\text{disc}(\mathcal{M})}.$$

Thus, for any $x \in \mathbb{R}$, it follows that

$$F_{\rho,b,\text{disc}(\mathcal{M})}(x) \leq F_{\text{adapted}(\rho,b),\mathcal{M}}(x) \leq F_{\rho,b,\text{disc}(\mathcal{M})}(x + H\eta)$$

where $F_{\rho,b,\text{disc}(\mathcal{M})}$ is the CDF of $Z_{\rho,b,\text{disc}(\mathcal{M})}$ and $F_{\text{adapted}(\rho,b),\text{disc}(\mathcal{M})}$ is the CDF of $Z_{\text{adapted}(\rho,b),\mathcal{M}}$.

Based on these arguments and Theorem H.3 in (Wang et al., 2023), we conclude:

$$\text{CVaR}_\tau(\text{adapted}(\rho, b); \mathcal{M}) \geq \text{CVaR}_\tau(\rho, b; \text{disc}(\mathcal{M})) - \tau^{-1}H\eta. \tag{11}$$

2) From $\mathcal{M}$ to disc($\mathcal{M}$): Let's introduce the memory-MDP model as defined in (Wang et al., 2023) first. The memory-MDP mode augments a standard MDP with a memory generator M_h, which produces memory items $m_h \sim M_h(s_h, a_h, r_h, \mathcal{H}_h)$ at each timestep. These memories are stored into the history $\mathcal{H}_h = (s_t, a_t, r_t, m_t)_{t\in[h-1]}$. The process of executing π in this memory-MDP is as follows: for any $h \in [H]$, $a_h \sim \pi_h(s_h, \mathcal{H}_h)$, $s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h)$, $r_h = r(s_h, a_h)$ and $m_h \sim M_h(s_h, a_h, r_h, \mathcal{H}_h)$. As a result of this process, the augmented MDP with memory has a history $\mathcal{H}_h^{\text{Aug}} = (s_t, b_t, a_t, r_t, m_t)_{t\in[h-1]}$. This memory-MDP model allows us to capture and model dependencies on past experiences through the memory items.

Building on the framework presented in (Wang et al., 2023), consider a scenario where we have a policy $\rho \in \Pi^{\text{Aug}}$ and an initial budget $b \in [0, 1]$, which we intend to use in the original MDP $\mathcal{M}$. To adapt this policy to run in disc($\mathcal{M}$), we introduce a discretized policy, which is history-dependent and incorporates memory. This policy operates in the discretized MDP disc($\mathcal{M}$) and is defined as follows:

$$\text{disc}(\rho, b)_h(s_h, m_{1:h-1}) = \rho_h(s_h, b - m_1 - \cdots - m_{h-1}).$$

Indeed, this definition of the discretized policy disc(ρ, b) is designed to ensure that, despite receiving discrete rewards $\hat{r_h}$ in the discretized MDP disc($\mathcal{M}$), the memory element m_h is carefully generated to imitate the reward that would have been received in the true MDP $\mathcal{M}$.

By applying Lemma H.2 in (Wang et al., 2023), we almost surely have

$$Z_{\rho,b,\mathcal{M}} \preceq Z_{\text{disc}(\rho,b),\text{disc}(\mathcal{M})}.$$

Consequently, if we define $F_{\rho,b,\mathcal{M}}$ as the CDF of $Z_{\rho,b,\mathcal{M}}$ and $F_{\text{disc}(\rho,b),\text{disc}(\mathcal{M})}$ as the CDF of $Z_{\text{disc}(\rho,b),\text{disc}(\mathcal{M})}$, we can establish that,

$$\forall x \in \mathbb{R} : \quad F_{\text{disc}(\rho,b),\text{disc}(\mathcal{M})} \leq F_{\rho,b,\mathcal{M}}.$$

Based on these observations and utilizing Theorem H.4 in (Wang et al., 2023), we obtain

$$\text{CVaR}_\tau^\star(\text{disc}(\mathcal{M})) \geq \text{CVaR}_\tau^\star(\mathcal{M}). \tag{12}$$

Combining Eq. (11) and Eq. (12), we have

$$|\text{CVaR}_\tau^{\rho^\star}(s_1; r) - \text{CVaR}_\tau^{\hat{\rho}}(s_1; r)| \leq \tau^{-1}H\eta. \tag{13}$$

We can satisfy the assumption about the optimization error by selecting $\eta \leq \epsilon\tau/3H$ to ensure

$$|\text{CVaR}_\tau^{\rho^\star}(s_1; r) - \text{CVaR}_\tau^{\hat{\rho}}(s_1; r)| \leq \epsilon/3.$$

C. Proof of Lower Bound

In this section, we prove our lower bound presented in Theorem 5.1. First, we develop the connection between the reward-free problem and the CVaR-reward-free RL problem.

Lemma C.1. *For any MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ with initial state s_1 and any policy π, there exists another MDP $\mathcal{M}' = (\mathcal{S}', \mathcal{A}, H+1, \mathbb{P}', r')$ with initial state s_0, we have*

$$\mathit{CVaR}_\tau^{\pi,\mathcal{M}'}(s_0) = \mathbb{E}_\pi\left[\sum_{h'=1}^{H} r_{h'}(s_{h'}, a_{h'})\middle| s_1, \mathcal{M}\right]. \tag{14}$$

Proof. We set horizon h starting at 0 in $\mathcal{M}'$. We can build such a $\mathcal{M}' = (\mathcal{S}', \mathcal{A}, H+1, \mathbb{P}', r')$, where $\mathcal{S}' = \mathcal{S} \cup s_0, s'$, $\mathbb{P}'(\cdot|s,a) = \mathbb{P}(\cdot|s,a)$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, $\mathbb{P}'(s_1|s_0,a) = \tau$ for any $a \in \mathcal{A}$, $\mathbb{P}'(s'|s_0,a) = 1-\tau$ for any $a \in \mathcal{A}$, $\mathbb{P}'(s'|s',a) = 1$ for any $a \in \mathcal{A}$, $r'(s,a) = r(s,a)$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, $r(s_0,a) = 0$ for any $a \in \mathcal{A}$, and $r(s_1,a) = 1$ for any $a \in \mathcal{A}$.

For any policy π, $\sum_{h'=1}^{H} r_{h'}(s_{h'}, a_{h'})$ equals to H with probability at least $1-\tau$. Thus, the τ-VaR following by any policy π in the transferred MDP $\mathcal{M}'$ is H. We have

$$\begin{aligned}
\mathrm{CVaR}_\tau^{\pi,\mathcal{M}'}(s_0) =& \max_{b_0 \in [0,H]} \{b_0 - \tau^{-1} V_0^{\pi,\mathcal{M}'}(s_0, b_0)\} \\
=& H - \tau^{-1}\mathbb{E}_\pi\left[\left(H - \sum_{h'=0}^{H} r'_{h'}(s_{h'}, a_{h'})\right)\middle| s_0, \mathcal{M}'\right] \\
=& H - \tau^{-1}\tau\mathbb{E}_\pi\left[\left(H - \sum_{h'=1}^{H} r'_{h'}(s_{h'}, a_{h'})\right)\middle| s_1, \mathcal{M}'\right] \\
& \underbrace{- \tau^{-1}(1-\tau)\mathbb{E}_\pi\left[\left(H - \sum_{h'=1}^{H} r'_{h'}(s_{h'}, a_{h'})\right)\middle| s', \mathcal{M}'\right]}_{=0} \\
=& \mathbb{E}_\pi\left[\sum_{h'=1}^{H} r'_{h'}(s_{h'}, a_{h'})\middle| s_1, \mathcal{M}'\right] \\
=& \mathbb{E}_\pi\left[\sum_{h'=1}^{H} r_{h'}(s_{h'}, a_{h'})| s_1, \mathcal{M}\right].
\end{aligned} \tag{15}$$

□

Now we can prove our lower bound, Theorem 5.1. Here, we restated Theorem 4.1 in (Jin et al., 2020), which show that any reward-free exploration algorithm that output ϵ-optimal policy must collect at least $\Omega(S^2AH^2/\tau\epsilon^2)$ trajectories in expectation.

Theorem C.2. *(Theorem 4.1 in (Jin et al., 2020)) Consider a universal constant $C > 0$. For a given risk tolerance $\tau \in (0,1]$, if the number of actions $A \geq 2$, the number of states $S \geq C\log_2 A$, the horizon $H \geq C\log_2 S$, and the accuracy parameter $\epsilon \leq \min\{1/4\tau, H/48\tau\}$, then any reward-free exploration algorithm that can output ϵ-optimal policies for an arbitrary number of adaptively chosen reward functions with a success probability $\delta = 1/2$ must collect at least $\Omega(S^2AH^2/\tau\epsilon^2)$ trajectories in expectation.*

Thus, any CVaR-RF exploration algorithm must collect at least $\Omega(S^2AH^2/\epsilon^2)$ trajectories from the state s_1, in expectation, and then collect at least $\Omega(S^2AH^2/\tau\epsilon^2)$ trajectories from the initial state s_0.

D. Technical Lemmas

D.1. An Essential Lemma for Upper Bound

The following crucial lemma establishes a relationship between the errors at step h and those at step $h+1$.

Lemma D.1. *On the event $\mathcal{E}$, for all $h \in [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$E_h^t(s,a) \leq 3H \left[\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \wedge 1 \right] + \sum_{s' \in \mathcal{S}} \mathbb{P}_h(s'|s,a) E_{h+1}(s', \rho^{t+1}(s')).$$

Proof. By the definition of $E_h^t(s,a)$ and the greedy policy ρ^{t+1}, if $n_h^t(s,a) > 0$,

$$E_h^t(s,a) \leq H \sqrt{\frac{2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \sum_{s' \in \mathcal{S}} \hat{\mathbb{P}}_h(s'|s,a) E_{h+1}(s', \rho^{t+1}(s')).$$

By the definition of $\mathcal{E}$ and Pinsker's inequality, we further have

$$\begin{aligned}
&\sum_{s' \in \mathcal{S}} \hat{\mathbb{P}}_h(s'|s,a) E_{h+1}^t(s', \rho^{t+1}(s')) \\
&\leq \sum_{s' \in \mathcal{S}} \mathbb{P}_h(s'|s,a) E_{h+1}^t(s', \rho^{t+1}(s')) + \sum_{s' \in \mathcal{S}} \left(\hat{\mathbb{P}}_h(s'|s,a) - \mathbb{P}_h^t(s'|s,a) \right) E_{h+1}^t(s', \rho^{t+1}(s')) \\
&\leq \sum_{s' \in \mathcal{S}} \mathbb{P}_h(s'|s,a) E_{h+1}^t(s', \rho^{t+1}(s')) + \|(\hat{\mathbb{P}}_h(\cdot|s,a) - \mathbb{P}_h^t(\cdot|s,a)\| \cdot H \\
&\leq \sum_{s' \in \mathcal{S}} \mathbb{P}_h(s'|s,a) E_{h+1}^t(s', \rho^{t+1}(s')) + H \sqrt{\frac{2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}},
\end{aligned}$$

where we use the fact that $E_{h+1}^t(s', \rho^{t+1}(s') \leq H$. Therefore, plugging in this inequality and using $2\sqrt{2} \leq 3$, we have

$$E_h^t(s,a) \leq \sum_{s' \in \mathcal{S}} \mathbb{P}_h(s'|s,a) E_{h+1}^t(s', \rho^{t+1}(s')) + 3H \sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}}.$$

Notice that

$$E_h^t(s,a) \leq H \leq 3H \leq 3H + \sum_{s' \in \mathcal{S}} \mathbb{P}_h(s'|s,a) E_{h+1}^t(s', \rho^{t+1}(s')),$$

and this is also true for $n_h^t(s,a) = 0$ with $1/0 = +\infty$, which leads to the conclusion. □

D.2. Auxiliary Lemmas

Lemma D.2. *VaR*$_\alpha = b^\star := \arg\max_{b \in \mathbb{R}} (b - \tau^{-1} \mathbb{E}[(b - X)^+])$.

Proof. Recall the definitions of CVaR and VaR, we have $\text{CVaR}_\tau(X) = \sup_b \left\{ b - \frac{1}{\tau} \mathbb{E}[(b-X)^+] \right\}$, $\text{VaR}_\tau(X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X \leq x) \geq \tau\}$. By Theorem 6.2 in (Acerbi & Tasche, 2002), we have

$$\text{CVaR}_\tau(X) = \mathbb{E}[X | X \geq \text{VaR}_\tau(X)].$$

Firstly, we define $f(b) = b - \frac{1}{\tau} \mathbb{E}[(b-X)^+]$, thus the derivative of $f(b)$ with respect to b is:

$$f'(b) = 1 - \frac{1}{\tau} \mathbb{P}(X \geq b).$$

By setting the derivative equal to zero, we have $\mathbb{P}(X \leq b) = 1 - \tau$. According to the definition of VaR, b is the τ-th quantile of the distribution of X, which means $b = \text{VaR}_\tau(X)$. Therefore, the critical point b^* that maximizes $f(b)$ is equal to $\text{VaR}_\tau(X)$. Now we prove $f(b^*) = \text{CVaR}_\tau(X)$.

$$\begin{aligned}
f(b^*) &= \text{VaR}_\tau(X) - \frac{1}{\tau} \mathbb{E}[(\text{VaR}_\tau(X) - X)^+] \\
&= \text{VaR}_\tau(X) - \frac{1}{\tau} \int_{-\infty}^{\text{VaR}_\tau(X)} (\text{VaR}_\tau(X) - x) dF(x) \\
&= \frac{1}{\tau} \int_{\text{VaR}_\tau(X)}^{\infty} x dF(x) = \mathbb{E}[X | X \geq \text{VaR}_\tau(X)] = \text{CVaR}_\tau(X).
\end{aligned}$$

□

Lemma D.3. *(Lemma F.1 in (Wang et al., 2023)) Given any $\rho \in \Pi^{Aug}$, $h \in [H]$, augmented state (s_h, b_h), and history $\mathcal{H}_h$, we have $V_h^{\rho}(s_h, b_h) = V_h^{\pi^{\rho,b}}(s_h, b_h; \mathcal{H}_h)$ for $b = b_h + r_1 + \ldots + r_{h-1}$. Particularly, $V_1^{\rho}(s_1, \cdot) = V_1^{\pi^{\rho,b}}(s_1, \cdot)$.*

Lemma D.4. *(Lemma 10 in (Kaufmann et al., 2021)) Given $\beta(n, \delta) = \log(2SAH/\delta) + (S-1)\log\left(e\left(1 + \frac{n}{S-1}\right)\right)$, it holds that $\mathbb{P}(\mathcal{E}) \geq 1 - \frac{\delta}{2}$. Furthermore, $\mathbb{P}(\mathcal{E}^{cnt}) \geq 1 - \frac{\delta}{2}$.*

Lemma D.5. *(Lemma 7 in (Kaufmann et al., 2021)) On the event $\mathcal{E}^{cnt}$, for all $h \in [H]$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$\forall t \in \mathbb{N}^*, \frac{\beta(n_h^t(s,a), \delta)}{n_h^t(s,a)} \wedge 1 \leq 4\frac{\beta(\bar{n}_h^t(s,a), \delta)}{\bar{n}_h^t(s,a) \vee 1}.$$

Lemma D.6. *(Lemma 19 in (Auer et al., 2008)) For any sequence of numbers $z_1, \ldots, z_n$ with $0 \leq z_k \leq Z_{k-1} = \max\left\{1, \sum_{i=1}^{k-1} z_i\right\}$,*

$$\sum_{k=1}^{n} \frac{z_k}{\sqrt{Z_{k-1}}} \leq (1 + \sqrt{2})\sqrt{Z_n}.$$

Lemma D.7. *(Lemma 15 in (Kaufmann et al., 2021).) Let $n \geq 1$ and $a, b, c, d > 0$. If $n\Delta^2 \leq a + b\log(c + dn)$ then*

$$n \leq \frac{1}{\Delta^2}\left[a + b\log\left(c + \frac{d}{\Delta^4}(a + b(\sqrt{c} + \sqrt{d}))^2\right)\right].$$